The Genetic Code for Proteins

Proteins consist of polypeptide chains which are formed by joining together amino acids head to tail. A typical chain may be some hundreds of residues long. Only 20 different amino acids are found in proteins no matter what the biological origin of the protein.

There are two sorts of nucleic acid, DNA and RNA. Each has a regular backbone consisting of an alternating sequence of phosphate and sugar groups. The sugar is ribose in RNA and the closely related deoxyribose in DNA. To each sugar is attached a flat heterocyclic molecule called a base. There are four sorts of bases in both DNA and RNA. In DNA they are adenine, guanine, cytosine and thymine. In RNA thymine (which is 5-mthyluracil) is replaced by uracil.

The genes cosist of <u>part</u> of a nucleic acid molecule. It is typically some hundreds of bases long. There is evidence that the amino acid sequence of any protein is determined by its own particular gene. Thus in some way a sequence of four different things (the bases in nucleic acid) determines the sequence of 20 different things (the amino acids in protein). The problem of how this is done is known as the coding problem.

There is in all cells an intricate biochemical mechanism for protein synthesis, parts of which are fairly well understood. In

brief, the synthesis takes place on particles about 200 Å in diameter called ribosomes. The genetic message is conveyed from the DNA of the gene to the ribosome by means of a special copy of the gene called messenger RNA. It is possible to break open cells and isolate the components necessary for protein synthesis and to combine them in such a way that a limited synthesis of protein takes place in the testtube. It is also possible to add artificial messenger RNA and to observe whidh amino acids are incorporated into protein as a result.

It has been shown by genetic experiments that in all probability three bases code one amino acid and that the message is read into the correct groups of three by starting at some fixed point, probably an end, and reading along three at a time. The work was done using the B gene of the r_{II} locus of bacteriophage T4 which attacks E. Coli. The mutants studied were the type produced by acridines which are believed to have one or more bases added or subtracted from the normal gene. Such mutants are quite distinct from those produced by chemicals which change one base into another. It was found that the acridine mutants could all be given either the sign + or the sign -, such that when joined into pairs by genetic recombination the pairs (++) or (--) were inactive, whereas many of the pairs (+-) gave an active or partly active gene. The striking result was that combinations of the type (+++) or (---) were also active. This suggested that three bases (or, less

likely, a multiple of three bases) stood for one amino acid.

Not all of the combinations (+-) were active, especially if they were rather far apart on the genetic map. A plausible explanation of this phenomenon made it likely that only a small number of the 64 possible triplets did stand for an amino acid, and that therefore the code was degenerate; that is, each amino acid would be represented by more than one triplet.

Recently biochemical work, especially by Nirenberg and Ochoa and their respective colleagues, has uncovered some of the details of the genetic code. Nirenberg and Matthaei found that if poly U (the RNA in which every base is uracil) is used as a messenger in the cell-free system it produced polyphenylalanine; that is a "protein" in which every amino acid residue is phenylalanine. This suggests that the triplet UUU stands for phenylalanine. More recent work using random co-polymers of two or more bases has suggested triplets for many of the other amino acids. There is good evidence that at least two triplets stand for the amino acid leucine and it is probable that each amino acid is represented by several triplets. However, the triplets coding one amino acid all appear to be rather similar. It is not yet known whether the 64 triplets can be put into 20 groups in some logical fashion or whether the groupings found are mainly due to historical accident.